Metadata schema

With MongoDB chosen as a solution to be used for storage of the data, the team came up with a metadata schema that we hope is employed by future tests, so that results from different samplings can all share the same standard and therefore be compared at will under whatever lens researchers choose. We will now present this standard and explain it.

```
{
   "_id": {"$oid": "00001"}
   "domain": "test.org",
   "url":
"https://www.test.org/signup",
   "rank": "1000",
   "testable": "Yes",
   "code": "<input
type="email">",
   "comments": "Field triggers
a captcha",
   "mailboxes":
   {
        "mail1": "Accepted",
        "mail2": "Accepted",
        "mail3": "Accepted",
        "mail4": "Rejected",
        "mail5": "Rejected",
        "mail6": "Rejected"
   }
}
```

These are the functionalities of each string:

- **_id**: Auto-generated unique identifier within the entire database.
- **domain**: Domain, not prefixed by “www”.
- **url**: URL containing the full path to the page containing the form.
- **rank**: Ranking of the website within that particular collection.
- **testable**: Indicates if the team ultimately managed to test the website.
- **code**: Contains the string that validates the form, if found.
- **comments**: Any relevant comment.
- **mailboxes**: List of ratings for each test case.

For future tests, we contemplate the inclusion of an “**eai**” field, that would indicate whether a given domain’s mail server supports Email Address Internationalization, which fundamentally means that they would be able to exchange mail in UTF-8 and

consequently in Unicode. This is a feature that has been gaining momentum in recent years, and it would be valuable to track its progress.

In the current study, 414 mailservers from websites were detected to have support for EAI.

For this particular 2019 Global Report, a "**country**" string had been included, so that a better sense of the geography of the list could be assessed. This, however, is not being suggested as part of the standard, due to the changes to the WHOIS database that no longer allows for massive lookups such as the one performed by the team, which would make it impractical moving forward. If RDAP eventually allows for such queries to be made, the string might be made viable again.